Bernáth László

Miért ne tulajdonítsunk semmilyen felelősséget tudattal nem rendelkező mesterséges intelligenciának?*

I. Felelősségtulajdonítás, mesterséges intelligencia és autonóm ágensek

A mesterséges intelligenciáról és autonóm ágensekről szóló filozófiai és nem filozófiai szakirodalomban egyre több olyan tanulmány jelenik meg, amelyek amellett érvelnek, hogy valamilyen, oksági felelősségen túlmutató, normatív tartalommal bíró felelősséget joggal tulajdoníthatunk olyan gépeknek vagy szoftvereknek, amelyek kellően nagy oksági autonómiával rendelkeznek (lásd többek között: Bechtel 1985; Dennett 1997; Floridi–Sanders 2004; Johnson–Powers 2005; Stahl 2006; Sullins 2006; Coeckelbergh 2009; Hage 2017). Általában azt nem szokás vitatni, hogy olyan gépeknek értelmetlen lenne felelősséget tulajdonítani, amelyek nagymértékben kézi vezérlés alatt állnak, vagy amelyek egy olyan egyszerű programot követnek, amely programot áttekintve viselkedésüket – viszonylag könnyen – teljesen vagy majdnem teljesen megmagyarázhatjuk. Nem merül fel, hogy bármiféle felelősséget tulajdonítsunk folyamatos beavatkozást igénylő drónoknak, egyszerű program által vezérelt robotporszívóknak, vagy éppen számítógépes játékokban a játékos ellenfeleit irányító mesterséges intelligenciának. Ugyanakkor egyre erősebbek azok a hangok, amelyek amellett érvelnek, hogy az ún. autonóm ágenseknek vagy gépeknek valamilyen normatív súllyal bíró felelősség tulajdonítható *attól függetlenül*, hogy ezeknek a gépeknek pontosan olyan elmét nem tulajdonítanánk, mint amilyen elmével az emberek rendelkeznek. Jelesül ezek az írások vagy explicite vagy implicite azt állítják, hogy nem számít, hogy ezek a gépek rendelkeznek-e *fenomenális tuda-*

* A tanulmány az Információs és Technológiai Minisztérium, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal által támogatott Posztdoktori Kiválósági Program (PD131998) keretében készült el. Emellett a K132911 (OTKA) pályázat támogatását élveztem. Külön köszönöm a két anonim bíráló segítő megjegyzéseit, ami alapján jelentős mértékben változtattam az eredeti szövegen.

tossággal vagy sem, joggal tulajdoníthatunk nekik felelősséget, amennyiben az egyéb szükséges kritériumokat is teljesítik.[1]

Mik is azok az autonóm ágensek? A szakirodalomban autonóm ágenseken[2] olyan mesterségesen létrehozott programokat vagy gépeket értenek, amelyek az oksági autonómia egy magasabb szintjével rendelkeznek, mint a korábbi programok és gépek. A nem autonóm gépek vagy programok vagy gyakori beavatkozást igényelnek az ember részéről ahhoz, hogy megoldják azt a feladatot, amelyet meg kell oldaniuk, vagy olyan programmal rendelkeznek, amelynek eredeti kódját ismerve viszonylag könnyen visszafejthető, miért éppen úgy próbálják megoldani a rájuk bízott feladatot, ahogyan. Az ilyen program vagy gép az őt irányító ember vagy a programozó meghosszabbított keze. Példaként felhozhatjuk azokat a sakkprogramokat, ahol bizonyos stratégiák vannak előre beprogramozva gépbe vagy egy kettesszintű „önvezető" autót, amely a cél eléréséhez mindvégig igényli a vezető felügyeletét és beavatkozását. Az ilyen rendszerekkel ellentétben az autonóm program, gép vagy ágens nem egyszerűen a programozó vagy az irányító meghosszabbított keze. Már csak azért sem, mert az autonóm ágenseket – hacsak nincs vészhelyzet vagy valamilyen rendkívüli esemény – nem tartják emberek a közvetlen irányításuk alatt. A programozó meghosszabbított kezének sem tekinthető a gép, mivel a kiinduló program egyik vagy másik elemével nem igazán lehet megmagyarázni, miért éppen azt a megoldást választotta a gép, amelyiket. Esetleg azért, mert a kiindulóprogram olyan óriási komplexitással bír, hogy ez valamiért lehetetlen, vagy azért – és valójában létező példát csak erre tudok mondani –, mert a gép mintegy maga alakította ki statisztikai tanulásra építő programjával azt a stratégiát, amellyel aztán a szóban forgó problémát megpróbálta megoldani. A statisztikai tanulás során a gép próba-szerencse alapon próbál megoldani egy-egy problémát, majd a visszajelzésekből (siker/kudarc) és az általa megfigyelt statisztikai összefüggésekből kikövetkezteti, hogy mi kell ahhoz, hogy a különböző helyzettípusokban az előtte álló feladatot sikeresen megoldja. Erre példák azok a modern sakkprogramok, amelyek már nagy biztonsággal képesek nagymestereket is elverni, mivel több millió sakkjátszma elemzésén és lejátszásán keresztül rendkívül bonyolult megoldáshálóval rendelkeznek arról, hogy milyen felállásnál milyen lépéseket érdemes tenni. Ugyanilyen autonóm ágensnek lehet tekinteni majd azokat az

[1] Érdemes megjegyezni, hogy Daniel Dennett egyfajta tudatosságot esszenciálisnak tart a morális felelősség szempontjából nézve, de ez a tudatosság nem fenomenális, hanem ahhoz kötött, hogy az adott gép vagy ember képes e másodrendű hiteket, vágyakat és egyéb mentális állapotokat kialakítani (hiteket a hiteiről, vágyakat a vágyairól, és így tovább). Dennett 1991, 1997.

[2] Jóllehet félrevezető (lásd Tollon 2019), én mégis – mivel a szakirodalomban megkövesedni látszik – használom az „autonóm ágensek" terminust. Miként a szakirodalom, én is csak annyit értek autonómia alatt, hogy (i) feladataikat emberi beavatkozás nélkül képesek (az esetek túlnyomó többségében) végrehajtani, és (ii) kiinduló programjuk ismerete önmagában nem segít hozzá tevékenységük részletes magyarázatához.

önvezető autókat, amelyek képesek lesznek a kiinduló ponttól a célig elvezetni anélkül, hogy az autóban ülő emberi „sofőr" ellenőrizné az autót.

Sokan úgy vélik, hogy ez a fajta autonómia – egyéb feltételek teljesülése esetén – elég alapot ad ahhoz, hogy az autonóm ágensek, az oksági felelősségen túl, valamilyen normatív súllyal rendelkező felelősséget is tulajdonítsanak. Azaz egyre többen érvelnek amellett, hogy ezek az autonóm ágensek nemcsak, hogy a fő és közvetlen okai annak, ha a rájuk bízott feladatot helyesen vagy helytelenül oldották meg, de olyan módon lehetnek fő és közvetlen okai a sikernek vagy valamilyen káreseménynek, hogy megfelelően járhatunk el, ha (morális vagy más normatív értelemben) dicsérjük, hibáztatjuk *(blame)* vagy büntetjük őket.[3] Például, ha egy önvezető autó a programozók által előre nem látható okokból hibás döntést hozva balesetet okoz, akkor nem a sofőrt, a programozót, vagy valaki mást kell morálisan hibáztatni, hanem az autót – feltéve, hogy további feltételek is teljesülnek –, valamint, ha a sakkprogram megveri a sakkmestert, akkor nem annyira a programozókat illeti dicséret, sokkal inkább a gépet.

Az egyéb feltételek közé a különböző szerzők általában olyasmiket sorolnak fel, amelyek képessé teszik a gépeket arra, hogy a felelősségre vonásra megfelelően reagáljanak. Tehát például valamiképpen képesnek kell lenniük információt szolgáltatni arról – ha nem is teljeset –, hogy miért tették azt, amit tettek, vagy legalább megfelelően reagálni azokra a gyakorlatokra, melyek jutalmazásként vagy büntetésként érik őket. Azaz a negatív visszajelzések hatására kevésbé legyenek hajlamosak újra elkövetni ugyanazt a hibát, a pozitív visszajelzések pedig erősítsék meg előnyös eljárásaikat. Mivel ezekhez a képességekhez megfelelő önreprezentációs képességekre van szükség, vagy legalábbis a kapott visszajelzésekre adott megfelelő reakciókra, ezért az autonóm ágenseknek való felelősségtulajdonítás kapcsán nem merül fel a fenomenális tudatosság jelentősége. Hiszen fenomenális tudatossággal akkor rendelkezik egy entitás, ha számára olykor valamilyen átélni azokat az eseményeket, melyek megtörténnek vele. Ám a fenomenális tudatosság nem szükségszerű feltétele annak, hogy egy gép rendelkezzen saját programjairól valamiféle reprezentációval, és annak sem, hogy legyenek olyan érzékelő és önszabályozó mechanizmusai, amelyeknek köszönhetően felismeri és helyesen reagál különböző emberi reakciókra. Mindehhez elegendő, ha a gép oksági profilja megfelelő szerkezetű, azaz a programjai által szabályozott folyamatai megfelelően kapcsolódnak egymáshoz.

[3] A dörzsölt olvasó kiszúrhatja, hogy itt a morális felelősségnek egy strawsoniánus meghatározását használtam. Azért tettem így, mert a szakirodalomban ez a legelterjedtebb formája a morális felelősség magyarázatának. Ugyanakkor Réz Anna egy érve (Réz kézirat), Szigeti András általános Strawson kritikája (Szigeti 2012), valamint Paár Tamással közösen írt tanulmányom gondolatmenete miatt (Bernáth–Paár, megjelenés előtt) elhagytam a strawsoniánus definíciót, amely a hibáztatáson és más reaktív attitűdökön keresztül definiálja a felelősséget. A gondolatmenet szempontjából nem lesz fontos, hogy a morális felelősséget így vagy máshogyan definiáljuk, mivel az a (személyes) morális megértés fogalmán fordul meg, amit semmilyen felelősség-meghatározásba nem célszerű beépíteni.

A következő szakaszban amellett fogok érvelni, hogy a morális értelemben vett felelősségtulajdonítástól hibás elválasztani a fenomenális tudatosság tulajdonítását. Fenomenális tudatosság nélkül ugyanis senki nem lehet felelős semmilyen tettéért, ráadásul az, hogy miért nem, viszonylag könnyen belátható, ha a helyes nézőpontot választjuk. Ezért ha valaki mégis morális felelősséget tulajdonít olyan gépeknek, amelyeknek nem tulajdonít egyszersmind fenomenális tudatosságot, annak az illetőnek tevékenysége inkoherens és végső soron értelmetlen.

II. MORÁLIS FELELŐSSÉG, MORÁLIS MEGÉRTÉS – MIÉRT NEM LEHETNEK MORÁLISAN FELELŐSEK A NEMTUDATOS AUTONÓM ÁGENSEK?

Azok a szerzők, akik szerint az autonóm ágensek számára akkor is joggal tulajdoníthatunk morális felelőséget, ha az autonóm ágenseknek nem tulajdonítunk egyszersmind fenomenális tudatosságot is, elsősorban abból a perspektívából közelítenek a problémához, hogy az autonóm ágensek morális hibáztatása éppen annyira hasznos lehet, mint az embereké, miközben egy autonóm ágens éppen olyan jól beilleszkedhet morális gyakorlatunkba, mint bárki más, mivel egy kellően fejlett autonóm ágens és egy ember működése között nincs érdemleges különbség. Ha finomszemcsés felbontásban nézzük az emberi viselkedést – legalábbis így szól az érv –, a fizika és a modern neurológia perspektívájából, akkor az ember viselkedését is éppen úgy determinisztikusan meghatározzák a bejövő információ hatására megtörténő neuronkisülések, mint ahogy egy autonóm ágens viselkedését meghatározza a kapott inputok hatására létrejött elektromos aktivitás a processzorban. Mindez persze azt is jelenti, hogy az a tény is irreleváns, hogy az embereknek fenomenális tudatosságot tulajdonítunk (nagy valószínűséggel veridikusan), míg az autonóm ágensek esetében ezt jóval kevésbé gondoljuk helyénvalónak. E különbség azért irreleváns e nézőpontból nézve, mert az oksági mélystruktúrában nem eredményez érdemi különbséget, hogy rendelkeznek-e az élőlények fenomenális tudatossággal vagy sem. Ám ha kevésbé közelítünk rá e két rendszer működésére, és úgymond hétköznapi életben felvett perspektívájából szemléljük őket, akkor az emberi döntéshozás és viselkedés egyaránt tekinthető olyan szabad és autonóm cselekvésnek, amelyet nem magyarázhatunk egyszerűen a körülmények kényszerítő erejével, hanem azt kell hogy mondjuk, azokat a (biológiai vagy mesterséges) ágensek valamilyen célra tekintettel, szabad akaratukból hozták meg (Bechtel 1985; Dennett 1997; Sullins 2006; Hage 2017).

Írásaik alapján az alábbi érvet lehet rekonstruálni az autonóm ágensek morális felelősségének elvi lehetősége mellett. Ezt az érvet Kiterjesztés-érvnek nevezem:

(a) Nincs olyan metafizikai különbség a nemtudatos magasan fejlett autonóm ágensek és az emberi lények között, amely különbség a morális felelősség vonatkozásában releváns lenne. (Nincs Különbség Tézis)
(b) Ha nincs olyan metafizikai különbség a nemtudatos magasan fejlett autonóm ágensek és az emberi lények között, amely a morális felelősség vonatkozásában releváns lenne, akkor az alapján kellene morális felelősséget tulajdonítanunk a nemtudatos magasan fejlett autonóm ágenseknek és az emberi lényeknek, hogy mennyire pozitívak e gyakorlat következményei. (Konzekvencialista Felelősség Tézis)
(c) Elvben kiderülhet, hogy összességében pozitív következményei lennének annak, ha morális felelősséget tulajdonítanánk nemtudatos magasan fejlett autonóm ágenseknek. (Lehetőség Tézis)
(d) Amennyiben a morális felelősség tulajdonítása valamely *T* típusú entitás számára összességében pozitív következményekkel jár, akkor *T* típusú létező morálisan felelős. (Revizionista Felelősség Tézis)

∴ Elvben kiderülhet, hogy a nemtudatos magasan fejlett autonóm ágensek morálisan felelősek.

Egy korábbi tanulmányomban (Bernáth 2021a) érveltem amellett, hogy ez az érv episztemológiailag sem nem jól megalapozott (a Nincs Különbség Tézis és a Lehetőség Tézis elégtelen alátámasztása miatt), sem nem koherens (mivel úgy támaszkodik etikai megfontolásokra, hogy közben a felhasznált etikai megfontolásoknál még alapvetőbb etikai megfontolásokat megbízhatatlannak bélyegez). Hiszen akár a fizika, akár a neurológia eredményeire támaszkodva nehéz megalapozni, hogy a fenomenális tudatosságnak nincs érdemleges és egyedi oksági szerepe az emberi viselkedés kialakításában,[4] és nehéz lenne megmondani, hogy hogyan tudnánk áttekinteni, hogy az autonóm ágenseknek morális felelősséget tulajdonító gyakorlat hatásai összességében pozitívak voltak-e vagy sem. Most azonban nem ezekre az érvekre helyezném a hangsúlyt, hanem a tanulmány egy másik, etikai gondolatmenetét hangszerelném át úgy, hogy világossá váljon, miért elhibázott az, ha valaki morális felelősséget tulajdonít egy nemtudatos gépnek, miközben nem tulajdonítana neki egyszersmind fenomenális tudatosságot is. Ugyanakkor ez az érvelés valójában a Kiterjesztés-érv mögötti egész funkcionalista alapállást támadja, amely szerint a morális gyakorlat elsődleges célja az lenne, hogy a társadalomban olyan viselkedésmintázatokat alakítson ki, melyek összességében pozitív következményeket eredményeznek. Ezért a Megértés-érv a (c) premisszán kívül az összes többi premisszáját támadja a Kiterjesztés-érvnek.

[4] Lásd például Mele 2009; Walter 2011; Shields 2014; Brass–Furstenberg–Mele 2019; Bernáth 2019, vagy magyarul Sági 2019; 2020)

A Megértés-érv abból indul ki, hogy az autonóm ágensek számára való morális felelősségtulajdonítás problémáját nem csak abból a perspektívából szemlélhetjük, hogy vajon hétköznapok során a magasan fejlett autonóm ágensek képesek lennének-e haszonnal és többé-kevésbé zökkenőmentesen beilleszkedni morális gyakorlatunkba, mint olyan entitások, melyek többnyire a morális szabályokhoz igazítják viselkedésüket, és akik – szükség esetén – morális reakciókkal (dicséret, kárhoztatás) e szabályok még következetesebb követésére késztethetőek. Onnan is nézhetjük, hogy mi a morális gyakorlat kitüntetett célja, amit csakis a moralitás személyes perspektívájából, annak személyes átéléséből ismerhetünk meg. Hogy megértsük, mi a baj a Kiterjesztés-érvvel, elsősorban arra van szükség, hogy ne engedjünk a kísértésnek, és a moralitásra ne mint a bölcs uralkodó egy eszközeként tekintsünk, hanem abból a perspektívából szemléljük, amely perspektívára mint a moralitás birodalmának polgárai teszünk szert hétköznapi életünkben, és akik e tapasztalatok fényében ismerik e birodalom határait, valamint e birodalom voltaképpeni célját.

Először érdemes arra fókuszálni, hogy mint a moralitás birodalmának tagjai ismerjük a konceptuális különbséget – és ezt a társadalom jólétéért aggódó filozófuskirályként sajnos hajlamosak vagyunk elfelejteni – *fegyelmezés* és *morális gyakorlat* (vagy rövidebben, de kevésbé pontosan: a moralitás) között. A fegyelmezés lehet a moralitás és a morális gyakorlat része, de a fegyelmezés önmagában nem feltétlenül morális jellegű. Ez egészen egyértelműen kiviláglik a magasabb rendű állatokkal vagy a kisgyermekekkel való kapcsolatunkban. Az állatokat és a kisgyermekeket sikeresen fegyelmezhetjük anélkül, hogy azt gondolnánk, egyszersmind morális gyakorlatot űzünk velük. A fegyelmezés olyan tevékenység, amelynek célja valamilyen stabil viselkedésmintázat kialakítása. Még ha a moralitásnak ez olykor célja is, akkor biztosan nem morális tevékenységről van szó, amikor egy kutyát arra próbálunk megtanítani, hogy ne ugráljon a vendégekre és a családtagokra.

Nemcsak, hogy a fegyelmezés nem mindig morális jellegű, de a morális gyakorlat sokszor nem foglal magában fegyelmezést, tehát egyáltalán nem célozza bizonyos viselkedési minta kialakítását. Nehéz azonban ennek bizonyítására példát találni, mert a legnagyobb bűnök büntetése általában igen félelmetes (úgymint a halálbüntetés vagy az életfogytiglani büntetés), és lehet érvelni amellett (jóllehet szerintem ez egyáltalán nem meggyőző), hogy valójában e legális és egyben morális gyakorlatoknak sem más a célja, mint a fegyelmezés, az elrettentés. Ám léteznek nagyon személyes morális gyakorlatok, amelyek bár egyfajta morális büntetést tartalmaznak, kizárhatjuk, hogy a fegyelmezés lenne a céljuk, mivel valamiért amúgy sem lenne mód már arra, hogy a szóban forgó gyakorlat bárkinek is a viselkedését megváltoztassa. Ezek többnyire súlyos és személyes ügyek. Mivel nem szeretnék saját példával élni, és sajnos megfelelő irodalmi példa sem jut eszembe, ezért az alábbi kitalált, de (már amennyire a filozófusok által kitalált történeteknél ez egyáltalán lehetséges) életszerű példát használom:

Apa és fiú

Egy apa szándékosan hazudik fiáról a fia szerelmének, hogy meghiúsítsa a számára nem kívánatos házasságot. A hazugság eléri célját, mert nagyon hihető volt, ráadásul egy olyan személytől származott, akinek látszólag semmi érdeke nem fűződött e hazugság kiagyalásához. A lány minden kapcsolatot megszakít a fiúval és többé nem hajlandó beszélni vele. A fiú számára mindvégig rejtve marad a szakítás valódi oka, de szerencsére később egy másik nő mellett megtalálja a boldogságát annak ellenére, hogy az apa folyamatosan belenyúl fia életébe, bár nem olyan drasztikusan és gonosz módon, mint a korábbi kapcsolat során. A fiú számára mindig sok kellemetlenség és szomorúság forrása volt, hogy apja nem hagyja azt, hogy tökéletesen a maga kezébe vegye az életét, de megalkuvásból és az apja iránti szeretetéből fakadóan megtanult ezzel együtt élni. Ám az apa idős korában súlyos beteg lesz, haldoklik a gyógyulás legkisebb reménye nélkül. A fiú ekkor tudja meg, mintegy véletlenségből, hogy mit tett apja hosszú évtizedekkel ezelőtt. Bemegy apja szobájába, és (amennyire erre képes) higgadtan megkérdezi apját, hogy emlékszik-e arra a lányra, és hogy mit tett vele. Az apja ennyi év távlatából – fia boldog házasságának évtizedei után – láthatóan nem érti, hogy fia miért hánytorgatja fel a múltat. Erre a fiú dühbe gurul, és keresetlen szavakkal elmagyarázza, hogy az események későbbi szerencsés fordulata után is miért fáj neki ennyire, amit megtudott a múltról, s miért volt apja tette minden egyébtől függetlenül rendkívül alávaló lépés, majd összefüggésbe hozza ezt az eseményt azzal, hogy apja egyfolytában igyekezett irányítást gyakorolni fia élete felett. Végül fogja magát, és kimegy a szobából, azzal az elhatározással, hogy többé nem akar beszélni az apjával.

Nos, azt hiszem, első látásra is világos, hogy ennek a viselkedésnek nem lehet az a célja, hogy egy új viselkedésmintát kialakítson az apában, mintegy megfegyelmezze, de az elrettentés is kizárt. Mikor tekinthetnénk sikeresnek a fiú morális gyakorlatát? Mit várnánk el a fiú helyében az apától? Azt hiszem azt, hogy az apa belássa, hogy mélyen rossz volt, amit tett (nemcsak a hazugság, hanem általában véve az irányításhoz való ragaszkodása és mindaz, ami ebből származott), és hogy belássa azt is, miért is volt rossz, amit tett, mindez megfelelően becsomagolva a lelkiismeret-furdalás vagy a szégyen érzésébe. Bár ez nem lenne így teljes: a teljes siker az volna, ha e morális belátás és a hozzá kapcsolódó morális érzések arra motiválnák az apát, hogy őszinte megbánással bocsánatot kérjen a fiától. Ez a példa szerintem jól mutatja a morális gyakorlatok elsődleges célját: olyan morális megértés létrehozása, amely összekapcsolódik a megfelelő morális érzelemmel és a lélek megfelelő átalakulásával.

Innen nézve az is érthető, hogy a puszta elrettentésben kimerülő fegyelmezés morális kontextusban miért tűnik üresnek. Van abban valami sivár és oda nem illő, ha a másikat a várható büntetéssel fenyegetjük, mivel a morálisan rossz

cselekvéssel az igazi probléma nem pusztán az, hogy a végrehajtója rosszul fog járni. Ugyanakkor – ha a büntetés eléggé biztosra vehető és valóban rettenetes – könnyen lehet, hogy egy ilyen fenyegető hangvételű fegyelmezés sikeres lehet abból a szempontból, hogy akár nagyobb hatásfokkal is rábírhatja az illetőt arra, hogy engedelmeskedjen a morálisan helyes cselekvés szabályainak. Az esetek többségében, azt hiszem, a sikkasztásra hajlamos fehérgalléros bűnözőket jobban visszatartja az, ha valaki kilátásba helyezi nekik a lebuktatásukat, mint ha valaki kifejezi, mennyire szégyenteljesnek tartja morális minőségüket.

Mindezzel egybevág, hogy hétköznapi morális gyakorlat során, amikor másokat morálisan kárhoztatunk vagy dicsérünk, tudatunk homlokterében nem az a vágy áll, hogy mások viselkedését megváltoztassuk. Leginkább úgy lehetne leírni, hogy milyen motiváció vezérel bennünket, amikor tényleg morálisan hibáztatunk másokat, hogy egy arra irányuló erős vágy vezérel minket, hogy az illető ismerje el, hogy igazunk van, érezze valamennyire rosszul magát azért, amit tett, és végül fejezze ki valamilyen formában a sajnálatát. Bármennyire is kínos ezt elismerni, de úgy vélem, hogy e hármas közül is prioritást élvez az első két komponens: az, hogy elérjük, hogy a másik felismerje, nekünk van igazunk, és a vételnek megfelelő tartalmú és intenzitású sajnálatot, szégyent vagy lelkiismeret-furdalást érezzen.[5] Ez magyarázza azt, hogy azt is szívesen hibáztatjuk, akiről úgy véljük, semmiképpen sem fogja kifejezni sajnálatát sem felénk, sem mások felé.

Fontos megjegyezni, hogy a szóban forgó morális megértés nem szenvtelen elfogadása valamilyen morális elvnek vagy állításnak. A korábbi példában az apa valószínűleg tudatában van annak, hogy hazudni más hitelének a megrontása érdekében morálisan rossz. Csak éppen alighanem úgy véli, hogy az ő speciális helyzetében a cél szentesítette az eszközt, és az ő konkrét esetében a fia boldogsága miatt fel volt hatalmazva arra, hogy hazudjon. A morális megértés abban áll, hogy mélyebben megértjük, hogy egy-egy morális szabály miért is rendelkezik kötelező erővel, akár nagyobb kötelező erővel, mint gondoltuk. Értékek és körülmények bonyolult viszonyát látjuk át jobban. Figyelmünket a morális szituációnak olyan aspektusai felé fordítjuk, aminek nem szenteltünk kellő figyelmet. Például az apa fia kirohanásának hatására jobban képes lehet arra, hogy az ő nézőpontjából nézze azt a helyzetet, amit valójában addig csak a saját szempontjából nézett. A fia számára az az első szerelem is fontos szerelem volt, s bár tudatában volt annak, hogy ama lehetséges házasságnak a boldogságát néhány külső körülmény is akadályozta volna, ennek ellenére ő akkor mégis e házasság megkötését látta jónak – és sokszor az önmagában nagyon nagy érték, hogy követhetjük saját megítélésünket arról, hogy adott helyzetben mi is lenne a jó. A fiú esetleg arra is rámutathatott kirohanásakor, hogy az apja aztán tudja,

[5] A morális hibáztatás *(blame)* ezen elemzését részletesebben alátámasztom egy másik tanulmányomban (Bernáth 2020).

hogy miben is áll a jósága annak, hogy úgy rendezhetjük az életünket, ahogyan akarjuk, hiszen az apa annyira rákapott erre, hogy mivel saját élete részeként tekint fia életére, azt is éppen olyan autonóm módon igyekezett igazgatni, mint a sajátját. *Belátni és átérezni*, hogy miért is értékes az önmeghatározás, és hogy *mennyire fájdalmas*, hogy mondjuk éppen az az önző vágyunk, hogy a saját önmeghatározásunk határait kitágítsuk, homályosította el annak észrevételét, hogy mekkora értéket veszünk el attól, akitől az autonómiáját vesszük el; nos, ilyesfajta *személyes*, belső felismerések azok, amelyeket *releváns morális megértésnek* tekinthetünk, ha magában a felelősségtulajdonításban a hétköznapi morális gyakorlat kitüntetett célját keressük.[6]

Az aligha vitatható, hogy a fenti értelemben vett személyes morális megértés teljességgel lehetetlen fenomenális tudat nélkül. Még ha egy nemtudatos gép ismeri is azt a szabályt, hogy törekedni kell a saját és mások autonómiájának a megőrzésére, nem értheti meg, miért jó (morálisan) e szabálynak a betartása, mivel csakis az autonóm cselekvés átélésén keresztül értheti meg bárki, *miért* érték az autonómia. Hasonlóképpen, elvben beprogramozhatjuk egy nemtudatos gépbe, hogy azoknak a cselekedeteknek a végrehajtására törekedjen, amelyek növelik az emberek boldogságát vagy jólétét, de mivel sohasem éli át, milyen az, amikor valaki viszonylag boldog vagy csak jól érzi magát, nem értheti, hogy e szabály követése miért jó. A nemtudatos gép már ezen az alapszinten képtelen a morális megértésre, nemhogy azon a magasabb szinten – aminek elérésére morális gyakorlatunk elsősorban törekszik –, amelynek lényege értékek egymáshoz való viszonyának megértése. Nem is beszélve arról, hogy e nemtudatos gépek nem élhetik át azokat a morális érzelmeket, amelyek képesek jelezni a morális tények súlyát, és amelyek elválaszthatatlan formai összetevői a személyes morális megismerésnek (úgymint a büszkeség, szégyen, lelkiismeret-furdalás, sajnálat stb.) Amire egy ilyen nemtudatos gép képes lehet, az legfeljebb az, hogy tanulóprogramját követve mind szofisztikáltabb szabályokat alakítson ki magának – de mindezt bármiféle morális megértés nélkül.[7]

[6] E koncepció sokat merít Fricker 2016-ból, de azt gondolom, jobban hangsúlyozza a morális megértés szerepét a kommunikatív elem rovására (egy korábbi tanulmányomban még nagyobb hűséggel követtem Miranda Fricker koncepcióját, lásd Bernáth 2021b).

[7] Számos olyan érvet megfogalmaztak már, amelyek nagyon hasonlítanak az itt ismertetettre. Ám ezek az érvek mind összekapcsolták a megértés problémáját azzal, hogy a gépek nem lehetnek képesek *kompetensen*, *megfelelő indokokra* támaszkodva morális döntéseket hozni (Johnson 2006; Himma 2009; Purves–Jenkins–Strawser 2015). Ez a fajta érvelés azonban nem elég erős az ellen a stratégia ellen, amikor a nemtudatos autonóm ágensek morális felelősségének a védelmezői azt hangsúlyozzák, hogy megfelelő programozottság esetén e gépek elvben képesek lehetnek arra, hogy morálisan ugyanolyan kívánatos döntéseket hozzanak, mint a sokszor irracionális és inkompetens emberek, így egyáltalán nem lehet azt mondani, hogy e gépek feltétlenül morálisan inkompetensek lennének vagy ne rendelkeznének a szükséges racionalitással. Floridi–Sanders 2004; Anderson–Anderson 2007 és Behdadi–Munthe 2020 javasolja e választ (utóbbiak tanulmányuk 205. oldalán), bár nem világos, mennyire tartják kielégítőnek). Hogy e dialektikai zsákutcát elkerüljem, a Megértés-érvben a hangsúlyt nem a

Ezért azoknál az autonóm ágenseknél, melyeket nem tartunk tudatosnak, legfeljebb annak van értelme, hogy fegyelmezzük őket anélkül, hogy morális felelősséget tulajdonítanánk nekik, ahhoz hasonlóan, mint ahogy az állatokat vagy a kisgyermekeket fegyelmezzük anélkül, hogy morális felelősséget tulajdonítanánk nekik. Ezek az autonóm ágensek a morális megértés teljes hiányában nem lehetnek a moralitás birodalmának polgárai, azaz morális ágensek, ennélfogva olyan morális tulajdonságokat sem lehet nekik tulajdonítani, mint a morális felelősség. Ez már csak onnan is belátható, hogy ha valaki vagy valami a legkevésbé sem értheti meg, hogy egyes cselekedeteknek miért az a morális státusza, ami, akkor morális értelemben nem lehet a tudatában annak, hogy mit is cselekszik.

Fontos megjegyezni, hogy a Megértés-érv nem arra a feltevésre támaszkodik, hogy valamilyen cselekvésért csak akkor viselhetünk morális felelősséget, ha azt a nagyon specifikus érzést átéltük már, ami közvetlenül kapcsolódik a szóban forgó cselekvés hatásaihoz. Például nem kell a Megértés-érv használójának azt az abszurditást is védelmébe vennie, hogy egy férfi nőgyógyász nem lehet morálisan felelős azért, ha műhibája megfoszt egy nőt az anyaság lehetőségétől, mivel neki nincs személyes tapasztalata az anyaságról. A Megértés-érv ugyanis pusztán arra a feltevésre támaszkodik, hogy a morális felelősséghez bizonyos alapvető fenomenális tapasztalatokra van szükség. Azaz nincs szükség arra, hogy a nőgyógyász átélje az anyaságot magát ahhoz, hogy megértse, miért problematikus az, ha hanyagság miatt elveszi valakitől ennek a lehetőségét. Elegendő hozzá az, hogy az illetőnek valamilyen alaptapasztalata legyen arról, hogy milyen fájó olykor elesni valamilyen lehetőségtől, hiszen ebből már megérti, hogy miért fontos odafigyelni és követni azokat a szabályokat, melyek csökkentik annak a valószínűségét, hogy embertársainknak jóvátehetetlen veszteségeket okozzunk.[8]

Kicsit formálisabban is megfogalmaznám azt az érvet, amely a fenti megfontolások alapján kirajzolódott.

Megértés-érv

(1) Azok az ágensek nem lehetnek morálisan felelősek, amelyek nem érthetik morális értelemben, mit cselekszenek, és/vagy a morális gyakorlatok célpontjaként nem segíthetik a morális gyakorlat kitüntetett céljának teljesülését.
(2) A morális gyakorlat kitüntetett célja a személyes morális megértés kialakítása.

morális kompetenciára és racionalitásra helyezem, hanem a morális teljesítménytől konceptuálisan (de nem feltétlenül okságilag) független (személyes) morális megértésre.

[8] E bekezdést az egyik anonim bíráló egy ellenvetéssel felérő megjegyzése miatt helyeztem el a szövegben. Köszönöm az anonim bírálónak, hogy felhívta a problémára a figyelmemet.

(3) Személyes morális megértés nélkül senki sem lehet tudatában annak, hogy morális értelemben mit is cselekszik.
(4) Azok az ágensek, amelyek nem rendelkeznek fenomenális tudatossággal, személyes morális megértéssel sem rendelkezhetnek.
(5) A nemtudatos autonóm ágensek *per definitionem* nem rendelkeznek fenomenális tudatossággal.

∴ A nemtudatos autonóm ágensek nem lehetnek morálisan felelősek.

Az érvet egyszerűbb formára is hozhatjuk:

Egyszerűsített Megértés-érv

(i) Azok az ágensek, akik/amik képtelenek személyes morális megértésre, nem rendelkezhetnek morális felelősséggel.
(ii) Azok az ágensek, akik/amik nem rendelkeznek fenomenális tudatossággal, azok képtelenek a személyes morális megértésre.
(iii) A nemtudatos autonóm ágensek *per definitionem* nem rendelkeznek fenomenális tudatossággal.

∴ A nemtudatos autonóm ágensek nem lehetnek morálisan felelősek.

A moralitás személyes perspektívájából nemcsak, hogy kirajzolódik a Megértés-érv, de kitűnik a Kiterjesztés-érv három premisszájának tarthatatlansága. Ugyanis innen nézve látható, hogy – szemben a Kiterjesztés-érv első premisszájával – a fenomenális tudatosság jelenléte vagy hiánya döntő morális és metafizikai különbség a nemtudatos autonóm ágensek és az emberek között. Ez még abban a furcsa esetben is így lenne, ha a személyes morális viselkedés nem lehetne érdemben hatással a viselkedésre, mert a személyes morális megértés kialakítása olyan önérték, amelynek kialakítása a morális gyakorlat kitüntetett célja. Az *Apa és fiú* című példából láthattuk, hogy egy morális gyakorlat akkor is sikeres lehet, ha a morális gyakorlat végül is senkit sem fegyelmez meg, és senkinek a viselkedését nem befolyásolja.

A Kiterjesztés-érv második és negyedik premisszáját is aláássa az, ha a moralitás személyes perspektíváját részesítjük előnyben. Ugyanis, ha ebből a perspektívából felismerjük, hogy a morális gyakorlat kitüntetett célja nem a viselkedésmintázat átalakítása, hanem a személyes morális megértés kialakítása, akkor áthidalhatatlan konceptuális különbség lesz a sikeres morális gyakorlat és a sikeres fegyelmezés között (utóbbinak valóban az a kitüntetett célja, hogy a viselkedést stabilan átalakítsa). Még ha valamilyen metafizikai vagy más vizsgálódás miatt kiderülne, hogy a morális gyakorlat kudarcra van ítélve (mert mondjuk le-

hetetlen személyes morális megértés), ez fogalmilag nem tenné lehetővé, hogy a moralitást és a fegyelmezési gyakorlatok egy részét fogalmilag azonosítsuk egymással, hiszen a kettő tulajdonképpeni célja egymástól különbözik – még ha igaz is az, hogy a morális gyakorlatokat használhatjuk és használjuk azzal a céllal is, hogy fegyelmezzenek. Ám mint ahogy egy labdát használhatunk arra is, hogy betörjünk vele egy ablakot, a labda játékszerből nem válik bontóeszközzé akkor sem, ha valamiért kiderül, valójában alkalmatlan arra, hogy örömöt csempésszen az életünkbe.

A személyes perspektíva felvétele végső soron leleplezi, hogy a nemtudatos autonóm ágenseknek való morális felelősségtulajdonítás ideája abból származik, hogy valaki összemossa a fegyelmezés gyakorlatát a morális gyakorlattal. A Kiterjesztés-érv híve nem tehet mást, ha ragaszkodni akar érvéhez, minthogy mégiscsak megpróbálja megmutatni, hogy a morális gyakorlat elsődleges, kitüntetett célja a fegyelmezés, azaz a viselkedésmintázatok átalakítása.

Azt hiszem, az erre irányuló érvek végül mind valamilyen evolucionista keretbe ágyazódnának be (hiszen kulturális vagy személyes tapasztalatra aligha hivatkozhatnának, mivel ezek a tapasztalatok szembe mennek ezzel az állítással). Eszerint csak látszat az, hogy lehet értelmes vagy sikeres valamilyen morális gyakorlatot folytatni úgy, hogy nincs igazi esély mások viselkedését befolyásolni (úgy, mint az *Apa és fiú* történet esetében). Ez a látszat abból ered, hogy az evolúció elrejtett bennünk egy motivációt, ami arra irányul, hogy lehetőség szerint minden gonosz tettet viszonozzuk a megfelelő morális reakcióval, mivel ez az egyszerű motiváció biztosítja leggazdaságosabban azt, hogy a morális gyakorlat összességében a lehető legnagyobb pozitív hatással lesz a viselkedésre.

Az ilyenfajta érvelés összekeveri azt, hogy valaminek mi az evolúciós oka és célja azzal, hogy valaminek mi a kitüntetett célja. A moralitáson kívül eső fegyelmezésen és a moralitás birodalmába eső fegyelmezés között éles különbséget tudunk tenni, mégpedig az alapján, hogy az utóbbinak sokkal nagyobb *normatív* súlya van, mint az előbbinek. Ha valaki bennünket próbál fegyelmezni, tehát valamilyen viselkedésre rábírni különböző jutalmazások és büntetések rendszerén keresztül, és e fegyelmezési kísérletet a moralitás területén kívülre helyezzük, akkor ezzel egyszersmind azt is mondjuk magukban, hogy ha szeretnénk és elég ügyesek vagyunk, akkor nyugodtan álljunk ellen az ilyen fegyelmezési kísérletnek. Nincs mélyebb normatív tétje a dolognak, minthogy elkerüljük-e a lehetséges büntetést vagy sem, és ha el tudjuk kerülni, akkor csináljuk csak azt, amiről mások megpróbálják elérni, hogy ne csináljuk. Egyszer például egy osztálytársam kitalálta – afféle dominanciajelzés gyanánt –, hogy ő márpedig napfénynek kitéve „érlelgetni” fogja a legmagasabb szekrény tetején a kapott ingyenes tejét, és mindenkit figyelmeztetett, hogy ne nyúljon hozzá, különben megbánja. Ha valaki erre azt gondolta volna, hogy ő csak azért is kiönti a megromlott tejet (aminek, hála a csomagolásnak, nem volt rossz szaga) úgy, hogy az erős srác nem veszi észre, azzal semmi gond nem lett volna, hiszen

a tej érlelése és a megőrzését kikiáltó „szabály" semmilyen mélyebb értékhez nem kapcsolódott. Az illető sem morálisan helyes, sem morálisan helytelen tettet nem követett volna el, és – amennyiben nem félt a büntetéstől – még azt is nyugodtan lerázhatta volna magáról, ha a srác észreveszi, hogy mit tett, és ezért számonkéri. Azaz *teljesen rendben van*, ha valaki ellenáll a fegyelmezésnek vagy figyelembe sem vesz egy puszta fegyelmezési kísérletet. Ezt jelenti, hogy nincs nagyobb normatív tétje a puszta fegyelmezésnek, jóllehet az *egzisztenciális tétje* ettől függetlenül igencsak nagy lehet, mivel az ellenállás vagy negligálás bizonyos helyzetekben akár életveszélyes megtorláshoz vezethet a fegyelmező részéről. Ezzel szemben a moralitás területére eső – fegyelmezési funkcióval is rendelkező – gyakorlatok nem ilyenek. Ha egy tolvajt a börtönbüntetésen és az ottani foglalkozásokon és fegyelmező eszközökön keresztül megpróbálják rávenni arra, hogy többé ne legyen tolvaj, *normatíve* elfogadhatatlan az, ha az illető vállat vonva lélekben továbbra is tolvaj marad – még akkor is, ha később már lehetősége sem lesz rá, hogy bárkitől bármit ellopjon. Még ha a moralitást az evolúció azért is alakította ki bennünk, hogy még hatékonyabbá váljanak fegyelmező eszközeink, a kultúra a moralitás eszköztárát azokhoz az értékekhez kötötte, melyeket igazán fontosnak tartunk, és a moralitás különleges súlyát (kialakulásától függetlenül) csakis a személyes tapasztalaton keresztül tudjuk megérteni. Csakis azért értjük, hogy a moralitás speciális *normatív* súllyal rendelkezik, és normatív súlya nem magyarázható pusztán azzal, hogy valakik bizonyos viselkedési mintákat akarnak bennünk és másokban kialakítani, mert fenomenális tudatosságunknak köszönhetően átéljük, hogy milyen kiemelt értéke van a boldogságnak, a szabadságnak, az igazságosságnak, az erényességnek és más hasonló dolgoknak szemben mondjuk azzal az értékkel, hogy valakinek a dominanciaigényét minden különösebb ok nélkül tiszteletben tartsuk. A moralitás speciális normatív státusza érthetetlen tehát a személyes perspektíva és a fenomenális tudatosság nélkül, így maga a moralitás mibenléte is, ezért a moralitás funkcióját is csak ezen keresztül érthetjük meg függetlenül attól, hogy mi volt a moralitás múltba vesző eredete.

III. A NEMTUDATOS AUTONÓM ÁGENSEK „GYENGÍTETT" FELELŐSSÉGE, A MEGÉRTÉS ÉRV ÉS A FIKCIONÁLIS BESZÉDMÓD

Sok szerző érzékeli, hogy jelentős problémákkal jár, ha teljes értékű morális felelősséget akarunk tulajdonítani olyan entitásoknak, melyeknek nem tulajdonítanánk egyszersmind fenomenális tudatosságot vagy valamilyen más – a morális gyakorlatunk alapján – morálisan releváns tulajdonságot. Ezért azt javasolják, hogy a felelősségnek valami más, normatíve kevésbé robusztus változatát tulajdonítsunk a nemtudatos autonóm ágenseknek (Floridi–Sanders 2004; Stahl 2006; Coeckelbergh 2009; Johnson and Powers 2005).

E tábor tagjai ugyan egymástól eltérő fogalmakat használnak, úgy vélem, e látszólagos sokféleség ellenére közös nevezőre lehet hozni a koncepciójukat. Felelősségre vonhatóságon (*accountability*, Floridi–Sanders 2004), szerepfelelősségen (Johnson and Powers 2005), kvázi-felelősségen (*quasi-responsibility* Stahl 2006), vagy virtuális felelősségen (*virtual responsibility*, Coeckelbergh 2009) olyasféle felelősséget értenek, amely ugyan nem azonos a valódi morális felelősséggel, de mégiscsak valamilyen felelősség, mivel a felelősségtulajdonítási gyakorlat haszna igazolhatja az autonóm ágensek vonatkozásában is azt a készтetésünket, hogy felelősséget tulajdonítsunk e gépeknek. Ráadásul – és ez az érv majd minden ilyen témájú cikkben megjelenik – mivel a morális felelősség empirikusan közvetlenül nem megfigyelhető feltételeinek teljesülésére vagy a fenomenális tulajdonság jelenlétére csak következtethetünk külső jegyekből, ezért e lefokozott felelősségfogalmak alkalmazása az autonóm gépekre lényegében ugyanazt a gyakorlatot teszi lehetővé, mint amit a morális felelősségtulajdonítás tesz lehetővé az emberek esetében azzal a különbséggel, hogy az utóbbiak esetében hiszünk az extra feltételek teljesülésében is, míg az előbbiek (az autonóm ágensek esetében) nem feltétlenül. Tehát e felelősségfogalmak lehetővé teszik, hogy a hétköznapokban kötelességeket tulajdonítsunk a gépeknek, és ha e kötelességeket nem teljesítik, akkor úgy tekintsünk ezekre a gépekre, mint akiket megfelelő módon hibáztathatunk és büntethetünk (különösen ha az átprogramozást egyfajta büntetésnek tekintjük és/vagy e gépek úgy vannak programozva, hogy hibáztató gyakorlataink hatására megváltozik a programjuk). Ugyanakkor e szerzők szerint nem kell egyszersmind abban is hinnünk, hogy e gyakorlatok segítségével szégyent, lelkiismeret-furdalást, morális megértést vagy tudást hozunk létre a gépben, mint ahogy abban sem, hogy a gépnek akárcsak esélye lenne érteni, miért is kellett volna betartania a szóban forgó kötelességet, azaz megfelelően működnie.

Kétféleképpen is lehet válaszolni ezen érvelésre. Egyrészt bárhogyan is tompítsuk a gépekre rótt kötelességek és felelősség normatív súlyát, valamekkora normatív súlya ezeknek megmarad, ám ekkor ugyanúgy alkalmazhatjuk a Megértés-érvet, mint a morális felelősség esetében. Ha a gépeknek nem tulajdonítunk fenomenális tudatosságot, akkor végső soron semmilyen tettnek nem érthetik semmilyen normatív súlyát vagy vonatkozását, és így nem lehetnek képesek arra, hogy bármilyen normatív erővel rendelkező kötelesség és felelősség birtokosai legyenek.

Másrészt válaszolhatunk egy olyan módon is, ami jobban figyelembe veszi ezeknek a szerzőknek a szándékait. Mint említettem, ezeket a szerzőket elsősorban az érdekli, hogy valahogy igazolják azt a jövőben valószínűleg megjelenő társadalmi gyakorlatot, amely alighanem haszonnal is fog járni. Nevezetesen, hogy úgy beszéljünk és bánjunk az autonóm ágensekkel, mintha azok többékevésbé ugyanúgy felelősek lennének cselekedeteikért, mint az emberek, jóllehet nem feltétlenül fogjuk ugyanakkor azt is elhinni róluk, hogy érezné-

nek bármit is vagy, hogy rendelkeznének szabad akarattal vagy valamilyen más olyan metafizikai jellemzővel, amiről jobbára azt gondoljuk, szükséges a *valódi* felelősséghez. Némileg szimpatizálok ezzel a fajta törekvéssel, ami sokkal szerényebb, mint a Kiterjesztés-érv képviselőié. Ám szerintem e szerényebb célkitűzést magukénak valló szerzőknek nem úgy kellene tenniük, mintha egy kisebb normatív súllyal rendelkező felelősségtípust próbálnának bevezetni, mert e felelősségtípusok – ha valódi felelősségtípusoknak tekintjük őket – ugyanúgy felvetik azt a kérdést, hogy hogyan is vehetnénk komolyan e felelősségtípusok tulajdonítását olyan esetekben, ahol nem tulajdonítanánk egyszersmind valódi megértést és (ezzel együtt) fenomenális tudatosságot is. Ehelyett azt kellene mondaniuk (mint ahogy a kvázi-felelősség és a virtuális felelősség terminusai tesznek is gesztust ez irányban), hogy valójában egy fikcionális beszédmód alkalmazását javasolják az autonóm gépek esetében, amelynek használata során azok terminusait nem érdemes szó szerint venni annak ellenére, hogy hasznos ez a fajta használat.[9]

A moralitás területén nagyon is használunk ilyen fikcionális beszédmódot. A kisgyermekekkel már akkor úgy beszélünk, mintha a moralitás birodalmának polgárai lennének, amikor még nem azok. Mindezt annak érdekében tesszük, hogy később azzá váljanak. Gyakran hasonlóan bánunk házi kedvenceinkkel is, bár ez esetben nem nevelő célzattal, hanem azért, mert ezzel is tovább antropomorfizáljuk őket, és úgy érezhetjük, hogy még közelebb állnak hozzánk. Fiktív karakterekről is gyakran úgy beszélünk, mintha felelősek lennének fiktív eseményekért, pedig jól tudjuk, hogy legfeljebb az író lehet felelős értük.

E moralitáshoz kapcsolódó fikcionális beszédmódnak éppen az a négy jellemzője van, amire azoknak van szükségük, akik valamilyen „felvizezett" felelősségfogalmat szeretnének tulajdonítani az autonóm gépeknek. Az első a hasznosság. A második, hogy e fikcionális beszédmódra olyannyira hajlamosak vagyunk, hogy könnyen beleéljük magunkat, mint egy jó regény olvasásába – ugyan valahol tudjuk, hogy szó szerint nem kell komolyan venni a mondottakat, mégis olyan tudatállapotba kerülünk, mintha komolyan vennénk (mintegy felfüggesztjük a hitetlenségünket). Ez jól illeszkedik ahhoz a tényhez, hogy

[9] Forrai Gábor nemrégiben terjesztett elő egy érvet, amely szerint, még ha ki is derülne, hogy az esetek túlnyomó többségében nem érdemlik meg az emberek, hogy morálisan hibáztassák őket a cselekedeteikért, akkor is nagyjából ugyanúgy kellene hagyni a már bejáratott morális gyakorlatunkat, ahogy van, mivel nagyon hatékony az a morális gyakorlat, amely az évezredek alatt mára kialakult (lásd Forrai 2021). Azt hiszem – bár emellett most nincs módom részletesebben érvelni –, hogy a fentiekhez hasonló okokból sokkal szerencsésebb lenne, ha a morális gyakorlatot egy hasznos fikciónak tekintenénk, és ilyen módon hagynánk többé-kevésbé érintetlenül hétköznapi megjelenési formáit, amennyiben kiderülne, hogy valójában az emberi lények sem felelősek soha vagy szinte soha cselekedeteikért. Ugyanis ekkor konzisztensen tudnánk a morális gyakorlat legdurvábban büntető gyakorlatait elhagyni (mivel a fikcionális beszédmódot nem vesszük a reflexió során teljesen komolyan), és megtartani a hétköznapi elemek nagy részét úgy, hogy a hétköznapokban nyugodtan beleélhetjük magunkat e fickionális morális világba.

minél fejlettebb szoftverrel és antropomorfabb külsővel rendelkeznek majd az autonóm ágensek, annál inkább hajlamosabbak leszünk beleélni magunkat abba a pillanat hatása alatt, hogy a gép éppúgy felelős az általa okozott eseményekért, mint mi. A harmadik, hogy még ha a pillanat hevében bele is süllyedünk a fikcionális beszédmódba, bizonyos határokon semmiképpen sem vinnénk túl a fikcionális beszédmód tartalmának komolyan vételét. Még ha a kutyánk antropomorfizálása érdekében úgy is beszélünk róla, mintha önálló erkölcsi cselekvő volna, amikor megharap valakit, egyértelmű, hogy a mi hibánk. Ha Darth Vaderről úgy is beszélünk, mintha többszörös személyiségzavarban szenvedne, nem gondoljuk komolyan, hogy egy húsvér pszichológus segíthetne rajta. Ehhez hasonlóan – úgy vélem – szerencsés lenne, ha a nemtudatos autonóm ágensek számára való felelősségtulajdonítást nem vennénk komolyan – és ez egybevág azoknak a szerzőknek az intencióival, akik valós, de gyengített felelősségfogalmakkal szeretnék a problémát kezelni. Azt ugyanis a szerzők többsége is elismeri, hogy problematikus lenne, ha az autonóm ágens károkozása során morális és legális gyakorlatunkban keresnénk elég lelkesen az emberi felelősöket, és nem vizsgálnánk a tervezők, programozók, illetve a felhasználók felelősségét. Ez a probléma azért rajzolódik ki különösen élesen, mert bár elméletben lehetséges a csoportos vagy megosztott felelősség, amennyiben egy cselekvőnek felelősséget tulajdonítunk, az valamelyest blokkolja vagy gyengíti az arra való hajlandóságunkat, hogy más ágenseknek is felelősséget tulajdonítsunk. Ezért ha valóban tulajdonítanánk valamiféle gyengített felelősséget az autonóm ágenseknek, akkor az várhatóan gyengítené azt a hajlandóságunkat, hogy az emberi felelősöket is megkeressük, ami összességében nem lenne túl kívánatos. Ezért előnyösnek tűnik, ha a gépeknek való felelősségtulajdonítási gyakorlatot fikciós gyakorlatnak tekintenénk. A negyedik, hogy bár szívesen használnánk a felelősségtulajdonításhoz kapcsolódó morális szótárat, a vele járó ontológiai elköteleződéseket nem feltétlenül vállalnánk fel. Még ha úgy is beszélünk Darth Vaderről, mintha számtalan ember haláláért lenne felelős egy messzi-messzi galaxisban, legtöbben nem vállalnánk fel az ezzel járó ontológiai elköteleződést még abban az értelemben sem, hogy posztuláljuk, létezik Darth Vader mint absztrakt entitás, és az író ezt az absztrakt entitást valahogy felruházta azzal az absztrakt tulajdonsággal, amely szerint „egy filmben meggyilkolt számtalan embert egy messzi-messzi galaxisban”.[10] Ehhez hasonlóan, legtöbben nem vállalnánk fel, hogy az autonóm ágensek teljesítik azokat a metafizikai feltételeket, melyek szükségesek a morális felelősséghez (például nem gondolnánk, hogy rendelkeznek fenomenális tudatossággal).

Összességében tehát úgy vélem, hogy mivel az autonóm ágensek a normativitás valódi megértésére egyáltalán nem képesek (sem morális, sem más te-

[10] Ugyanakkor léteznek olyan metafizikai elméletek, amelyek éppen ezt állítják, lásd például Zvolenszky 2012.

rületen), s mivel teljesen kielégítő lenne minden szempontból, ha az autonóm ágenseknek való felelősségtulajdonítást fikcionális beszédmódnak tekintenénk, ezért felesleges feltalálni vagy bevezetni olyan felelősségfogalmakat, amelyek gyengített, de valós felelősséget tulajdonítanak a nemtudatos autonóm ágenseknek.

IV. KONKLÚZIÓ HELYETT

Azt hiszem, mivel e tanulmány érvelése viszonylag egyszerű volt, nincs szükség arra, hogy ehelyütt még egyszer részletesen összefoglaljam. Inkább csak tisztázni szeretnék valamit a tanulmány végkövetkeztetésével kapcsolatban. A tanulmány csak amellett érvelt, hogy nemtudatos autonóm ágenseknek nincs értelme morális vagy nemmorális, normatív tartalommal rendelkező felelősséget tulajdonítani (az oksági felelősségen túl). Pontosabban: olyan autonóm ágenseknek nem szabadna normatív tartalommal bíró felelősséget tulajdonítani, amelyeknek nem tulajdonítunk, vagy nem kellene tulajdonítanunk egyszersmind fenomenális tudatosságot is. Ugyanakkor mindaz, amit mondtam, nem vonatkozik azokra az autonóm ágensekre, legyenek azok akár gépek, akár programok, amelyeknek jó okkal tulajdoníthatnánk fenomenális tudatosságot. Ilyen esetekben teljesen értelmes lehet az – legalábbis e tanulmányban nincsen semmi, ami ez ellen érvként lenne felhozható –, ha nemcsak fenomenális tudatosságot, de akár morális felelősséget is tulajdonítunk nekik.

Arra persze nincs ehelyütt módom (sőt elegendő kompetenciával sem rendelkezem hozzá), hogy részletezzem, milyen esetekben vagyunk hajlamosak egy entitásnak fenomenális minőséget tulajdonítani. Azt sem tudom kizárni 100%-os biztonsággal, hogy azok a gépek, amelyek körülvesznek minket (például sakkprogramok) nem rendelkeznek fenomenális tudatossággal. Így azt sem tudom kizárni, hogy már most sokan vannak, akik – helyesen – fenomenális tulajdonságokat és – akár szintén helyesen – morális felelősséget tulajdonítanak a sakkprogramoknak. Ugyanakkor mind a szakirodalomban, mind – úgy vélem – a „laikusok" között elterjedt, és úgy vélem, megfelelő alapokkal rendelkező meggyőződés az, hogy a már jól ismert autonóm ágensek nem rendelkeznek fenomenális tudatossággal, és nincs is jó okunk ilyet tulajdonítani nekik. Csak ha ez a két meggyőződés valóban jól megalapozott, van jelentősége e tanulmány gondolatmenetének, de úgy gondolom, joggal bízhatom abban, hogy e két feltevés valóban jól megalapozott, még akkor is, ha nem áll módomban azt kifejteni, hogy miért.

Mindezen megszorításokkal együtt kiemelt fontosságúnak tartom azt, hogy ne vegyük komolyan azt a késztetésünket, hogy felelősséget tulajdonítsunk olyan gépeknek, melyeknek nem tulajdonítanánk egyszersmind fenomenális tudatosságot is, és ne hallgassunk azokra a filozófusokra, akik ezt a könnyel-

műséget bátorítják. Egyrészt azért, mert jó ideig hétköznapjainkban azok az autonóm ágensek lesznek túlnyomó többségben, amelyeknek nem tulajdonítanánk fenomenális tudatosságot, és társadalmilag is meghatározó jelentőségű lehet, hogy helyesen vagy helytelenül viszonyulunk-e ezekhez az entitásokhoz. Másrészt pedig azért, mert nagyon fontos lenne nem elfelejtenünk azt, hogy a moralitás rendszere nem csupán és nem is elsődlegesen azt célozza, hogy a társadalom tagjait egyszerűen megfegyelmezze. A morális megértésnek önértéke van, a moralitás ezeknek az értékeknek a létrejöttét célozza kitüntetett módon, s szubjektív nézőpont és fenomenális tudatosság nélkül ez a fajta személyes morális megértés lehetetlen lenne.

IRODALOM

Bernáth, László 2021a. Can Autonomous Agents Without Phenomenal Consciousness Be Morally Responsible? *Philosophy & Technology*. 34/4. 1363–1382.

Bernáth, László 2021b. Defending libertarianism through rethinking responsibility for consequences. *Philosophical Papers*. 50/1–2. 81–108.

Bernáth, László 2020. Blame and Fault: Toward a Conative Theory of Blame. *Disputatio: International Journal of Philosophy*. 12. 371–394.

Bernáth, László 2019. Why Libet-style experiments cannot refute all forms of libertarianism. In Bernard Feltz – Marcus Missal – Andrew Sims (szerk.) *Free will, causality, and neuroscience.* Leiden–Boston, Brill–Rodopi. 97–119.

Bernáth, László – Paár, Tamás (megjelenés alatt). Responsibility first: How to Resist Agnosticism about Moral Responsibility. *Dialectica*.

Anderson, Michael – Anderson, Susan Leigh 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*. 28/4, 15–26.

Bechtel, William 1985. Attributing responsibility to computer systems. *Metaphilosophy*. 16/4. 296–306.

Behdadi, Dorna – Munthe, Christian 2020. A normative approach to artificial moral agency. *Minds and Machines*. 30. 195–218.

Brass, Marcel – Ariel Furstenberg – Alfred R. Mele 2019. Why neuroscience does not disprove free will. *Neurosci Biobehav Rev*. 102. 251–63.

Coeckelbergh, Mark 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & society*. 24/2. 181–189.

Dennett, Daniel C. 1997. When HAL Kills, Who's to Blame? In David G. Storck (szerk.) *HAL's Legacy: 2001's Computer as Dream and Reality*. Boston/MA, MIT Press. 351–366.

Dennett, D. C. (1991). *Consciousness Explained*. Boston/MA: Little, Brown and Company.

Floridi, Luciano – Sanders, Jeff W. 2004. On the Morality of Artificial Agents. *Minds and Machines*. 14. 349–379.

Forrai Gábor 2021. A felelősségtulajdonítás haszna, avagy miért téves a revizionista érv. *Magyar Filozófiai Szemle*. 65/1. 5–25.

Fricker, Miranda 2016. What's the point of blame? A paradigm based explanation. *Noûs*. 50/1. 165–183.

Hage, Jaap 2017. Theoretical foundations for the responsibility of autonomous agents. *Artif Intell Law*. 25. 255–271.

Himma, Kenneth Einar 2009. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*. 11/1. 19–29.

Johnson, Deborah G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*. 8/4. 195–204.

Mele, Alfred R. 2009. *Effective Intentions: The Power of Conscious Will*. Oxford, Oxford University Press.

Purves, Duncan – Jenkins, Ryan – Strawser, Bradley J. 2015. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*. 18/4. 851–872.

Réz Anna (kézirat): The Fairness of Holding Responsible: A Victim-Centered Approach.

Sági Péter Tamás 2020. Van szabadság, csak akarni kell. *Magyar Filozófiai Szemle*. 64/2. 5–25.

Sági Péter Tamás 2019. A szabadság idegtudománya. In Ruszkai Szilvia Éva – Furtado Renátó – Szabados Bettina – Szabó Ferenc (szerk.) *Ütközéspontok V. A Doktoranduszok Országos Szövetsége Filozófiatudományi Osztálya konferenciájának kötete*. Szeged, Jate Press. 117–127.

Shields, Grant S. 2014. Neuroscience and conscious causation: Has neuroscience shown that we cannot control our own actions? *Review of Philosophy and Psychology*. 5/4. 565–582.

Stahl, Bernd Carsten 2006. Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics Inf Technol*. 8. 205–213.

Szigeti, András 2012. Revisiting Strawson's Arguments from Inescapability. *Philosophica*. 85, 91–121.

Sullins, John P. 2006. When is a robot a moral agent. *International Review of Information Ethics*. 6/12. 23–30.

Tollon, Fabio 2019. Moral agents or mindless machines? A critical appraisal of agency in artificial systems. *Hungarian Philosophical Review / Magyar Filozófiai Szemle*. 63/4. 9–23.

Walter, Henrik 2011. Contributions of Neuroscience to the Free Will Debate: From random movement to intelligible action. In Robert Kane (szerk.) *The Oxford Handbook of Free Will*. 2. kiad. Oxford, Oxford University Press.

Zvolenszky, Zsófia 2012. Against Sainsbury's Irrealism About Fictional Characters. *Hungarian Philosophical Review / Magyar Filozófiai Szemle*. 56/4. 83–109.